A Machine Learning Approach to Estimate Surface-Level NO_2 Concentration using High Resolution Remote Sensing Observations

Aadi Kenchammana
Saint Francis High School
Mountain View, USA
Aadikenchammana@gmail.com

Theodore Trefonides
Spatiotemporal Innovation Center
George Mason University
Fairfax, USA
ttrefoni@gmu.edu

Chaowei Yang
Spatiotemporal Innovation Center
George Mason University
Fairfax, USA
cyang3@gmu.edu

***Abstract*—Current measurement systems for Nitrogen Dioxide (NO_2), a key contributor to poor air quality, remain inadequate. In situ networks provide precise surface-level readings but lack high-resolution spatial data. In contrast, satellite-based remote sensing provides wider coverage but doesn't provide surface-level measurements. Previous studies have shown that machine learning models built using remote sensing data can predict surface-level concentrations with some degree of accuracy. Here, we present machine learning models trained with NASA's TEMPO instrument which provides high-resolution data. Our work demonstrates that the best model trained with TEMPO data provided accurate hourly surface-level NO_2 concentrations and significantly outperformed previous models.**

***Index Terms*—air quality monitoring, machine learning, remote sensing**

I. Introduction

The World Health Organization estimates that polluted air is responsible for up to seven million deaths annually [1]. Nitrogen dioxide (NO_2) is a key pollutant that contributes to poor air quality [2]. Despite the well-documented harmful effects of NO_2, there are substantial limitations in the systems currently in place to measure this pollutant. There are two primary monitoring techniques:

1) In situ monitors: In situ monitoring networks consist of ground-based measurement stations that measure air pollutants at specific locations. These stations provide highly accurate and precise surface-level readings of NO_2 concentrations but are expensive and are often limited to urban areas [3]. This limitation means that rural and remote areas remain unmonitored, leading to gaps in data coverage and an incomplete understanding of regional air quality.
2) Satellite-based remote sensing: Satellite-based remote sensing offers an alternate approach to in situ monitoring by providing broader spatial coverage and finer spatial resolution. Satellites equipped with spectrometers can detect tropospheric NO_2 over large geographic areas, making it possible to monitor air quality cheaply on a global scale. However, these systems have their own set of limitations. Satellites can only measure the vertical columnar density (VCD) of NO_2 rather than surface-level concentrations [3]. VCD is the amount of NO_2 which is contained in a column of the atmosphere stretching from the ground to space. This is less actionable as only surface-level NO_2 affects air quality.

The inadequacies of current measurement systems highlight the need for improved technologies and methodologies to estimate surface-level NO_2 more effectively. Machine learning models built using remote sensing data have the potential of providing more accurate and comprehensive assessments of NO_2 concentrations that affect humans.

Previous studies using Sentinel-5 Precursor's Tropospheric Monitoring Instrument (TROPOMI) have demonstrated that remote sensing data, when combined with machine learning models, can predict surface-level pollutant levels with some degree of accuracy. The TROPOMI sensor provides daily observations of NO_2 VCD with a spatial resolution of 7 km x 7 km [4]. A recent study utilized TROPOMI satellite observations along with other meteorological parameters to train a neural network model for estimating daily surface NO_2 concentrations [5]. This model was validated against ground-based in situ air quality measurements and showed a relatively high coefficient of correlation (R) of 0.80, which equates to an R^2 value of 0.64. Another study focusing on the Alpine Plain trained an XGBoost model to predict hourly NO_2 using TROPOMI and achieved an R^2 value of 0.59 [6]. These efforts have demonstrated the potential of combining satellite data with machine learning to fill gaps left by traditional monitoring methods. However, models based on TROPOMI data are constrained by TROPOMI's temporal resolution, which is limited to a single daily measurement. This limitation hinders their ability to capture diurnal variations in surface-level NO_2 levels that are critical for understanding pollution dynamics in areas where NO_2 concentrations can fluctuate significantly throughout the day [7].

NASA's newly launched Tropospheric Emissions: Monitoring of Pollution (TEMPO) instrument addresses these limitations by measuring tropospheric pollutants at an unprecedented spatial (2.0 km x 2.0 km) and temporal (hourly) resolution [8].

The satellite was launched on April 7, 2023, and data from the sensor has only become recently available.

In this paper, we leverage TEMPO data to train machine learning models that can accurately predict surface-level NO_2 concentrations on an hourly basis and validate the results using in situ ground-based measurements.

II. Methods

In this section, we will provide details of our data preparation, model training, and validation.

A. Data Preparation

The data used to train the models included satellite NO_2 VCD observations, meteorological data, elevation, population density, and traffic density. They were processed as follows:

- Satellite-based NO_2 VCD measurements were obtained from both the TROPOMI and TEMPO sensors. We utilized L2 data from TROPOMI satellite data [4] which we converted to L3 using the HARP package [9] and further processed to fit into an hourly 1 x 1 km grid. The original data had a resolution 7 x 7 km and was re-gridded using the HARP package [9] to a resolution of 0.03 x 0.03 degrees and then projected onto a 1 x 1 km grid using bilinear interpolation. As TROPOMI provides daily measurements [4], we utilized linear time interpolation to scale the data to an hourly resolution. The TEMPO data [8] was similarly spatially re-gridded from a 2 x 2 km resolution to a 1 x 1 km resolution using bilinear interpolation.
- Meteorological data were acquired from the European Centre for Medium-Range Weather Forecasts' (ECMWF) ERA5 reanalysis product [10] at a 0.25 x 0.25 degree spatial resolution and hourly temporal resolution which we projected onto a 1 x 1 km grid using bilinear interpolation.
- Human population density was derived from the GlobPOP dataset [11], averaged over 2018-2020, and projected onto the same 1 x 1 km grid. This data was treated as a constant.
- Elevation data was obtained using a Digital Elevation model from United States Geological Survey (USGS) [12] was downloaded at a resolution of 30 x 30 m and projected onto a 1 x 1 km grid using bilinear interpolation. This data was also treated as a constant
- Traffic density was estimated using interstate highway traffic data from California's Department of Transportation [13] that we processed into a 1 x 1 km raster. This data only includes traffic density on interstate highways, but we utilized it as a proxy for traffic density in all roads in the surrounding area. This data was treated as constant as it has a yearly temporal resolution. We also got road density data at an 8 x 8 km resolution from the Global Roads Inventory Project (GRIP) dataset [14].

The regridded and rescaled data from these sources were matched to one another and compiled into a dataset with 24 covariates which was then used to train the candidate models. Any timestamp where there was less than 20% data coverage from TEMPO or TROPOMI was discarded from the data.

B. Model Training and Validation

TABLE I
HYPERPARAMETER TUNING RANGES FOR TREE-BASED MODELS

hyperparameter	Low Range	High Range
num_estimators	100	700
learning_rate	0.001	0.5
subsample	0.2	1
max_depth	1	11
cosample_bytree	0.1	1.0
gamma	0	0.4
reg_alpha	0	100
reg_lambda	0.001	10

Since we expected the relationships between the input features and the target feature (surface-level NO2) to be non-linear, we trained tree-based models and a Multi-Layer Perceptron (MLP) neural network model. The tree-based models were built with XGBoost and Random Forest. For each candidate model type, we ran three training runs: with only TEMPO data, with only TROPOMI data, and with both TEMPO and TROPOMI data. For both tree-based model architectures we ran a hyperparameter sweep (Table 1) to find the optimal model parameters which were used for all further training and testing. To do this, we utilized the GridSearchCV parameter sweep with the eight hyperparameters. For the MLP model, we constructed a 4-layer architecture with 64, 32, 16, and 1 node in the successive layers.

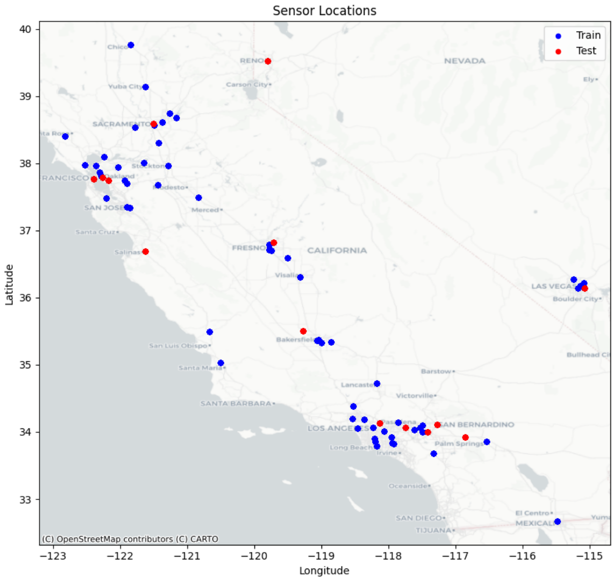

Fig. 1. Clusters of Sensor Locations

978-1-6654-7345-3/22/$31.00 ©2024 IEEE

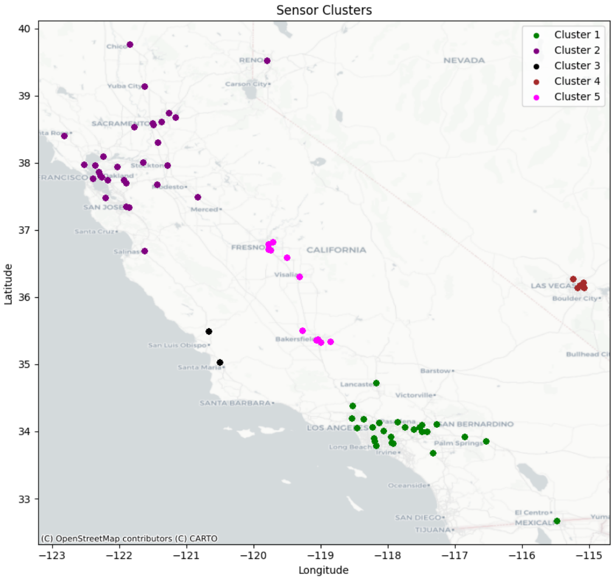

Fig. 2. Location of Training and Testing Sites

To reliably achieve a roughly equal geospatial distribution of the training and testing data points, we leveraged a spatially aware train/test split methodology. First, the 75 air quality monitoring sites were grouped into 5 clusters based on their spatial proximity using a K-means clustering approach. The clustered sensor locations are shown in Figure 1. For each of these clusters, 60% of the sensors (45 sites) were randomly chosen to be denoted as training sites while the remaining 40% of the sensors (30 sites) were denoted as testing sites (Figure 2). This novel split methodology ensured that the training and testing sites were equally distributed. Equal distribution of sensors ensures that the evaluation results reliably reflect the model's performance. Additionally, this allowed us to evaluate the model's generalizability as the model was evaluated on sensor locations which it wasn't trained on.

Our models were validated using surface-level NO_2 concentration data obtained from in situ monitoring networks. Surface-level NO_2 data were sourced from the EPA's AirNow service [15], which is considered the standard in air quality monitoring, providing high-accuracy sensor readings across the U.S. These measurements served as the ground truth for model training.

III. Results

Table 2 provides the testing R^2 for different model types and input data combinations we tested. The overall best performing model was an XGBoost model trained with TEMPO data, and it achieved an R^2 value of 0.73. This exceeded the previous best performing hourly model, which was trained using TROPOMI data, that achieved an R^2 value of 0.59. The two tree-based models performed better than the MLP model which we suspect is due to the relatively small amount of TEMPO data as it is a relatively new data source. The two months of data that is available is likely not enough for a complex deep learning model to converge on a solution.

TABLE II
Testing R^2 values

Model Types	TEMPO	TEMPO + TROPOMI	TROPOMI
XGBoost	0.73	0.72	0.59
RF	0.53	0.51	0.41
MLP	0.43	0.46	0.38

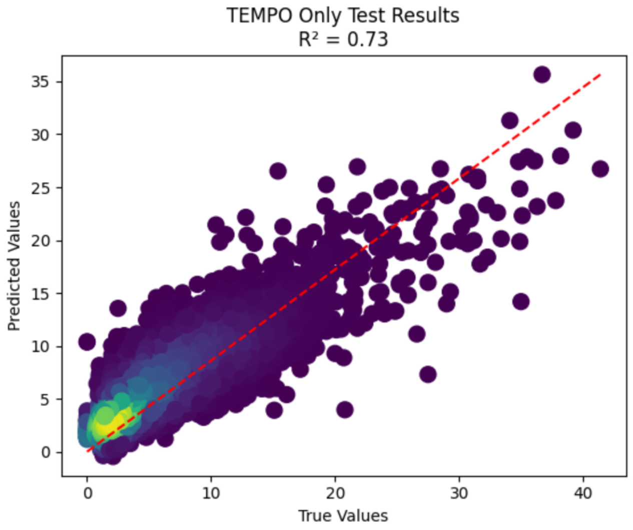

Fig. 3. Predicted vs actual NO_2 Concentration using XGBoost with TEMPO Data

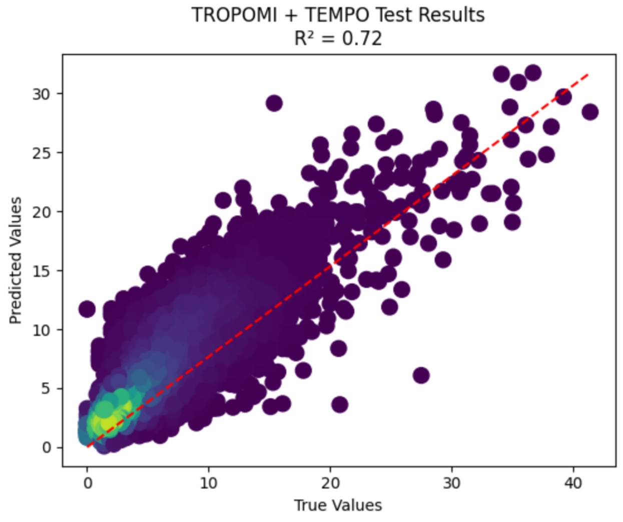

Fig. 4. Predicted vs actual NO_2 Concentration using XGBoost with TEMPO and TROPOMI Data.

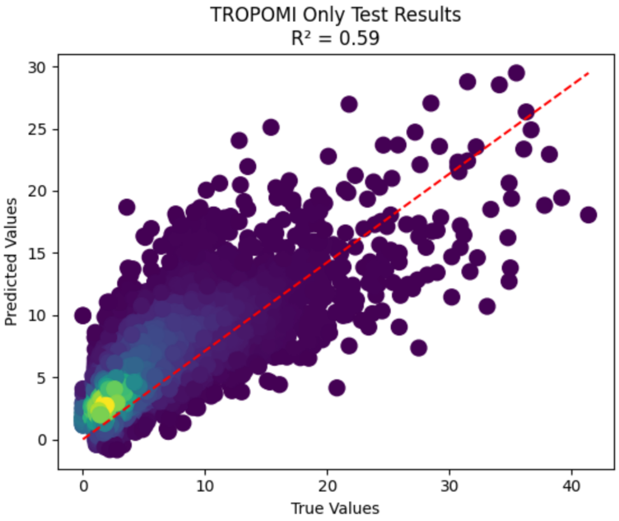

Fig. 5. Predicted vs actual NO_2 Concentration using XGBoost with TROPOMI Data.

The models trained with TEMPO data perform significantly better than the models trained with TROPOMI data. Both tree-based models performed the best when trained with TEMPO data only. The margin of improvement is substantial, varying between an increase in R^2 of 0.05 and 0.13, which suggests that the improved temporal and spatial resolution of the TEMPO satellite contributes to a meaningful improvement in model performance. The correlation plots in Figures 3-5 further illustrate the improved performance with TEMPO data. The cluster plot of the TEMPO-XGBoost model's evaluation shows a higher degree of linearity, indicating a stronger correlation (Figure 3). In contrast, the TROPOMI-XGBoost model's evaluation displays greater dispersion from linearity, reflecting a weaker correlation (Figure 5). This improved performance is further evident in Figure 6-7 which show that the XGBoost model trained with TEMPO data captures the diurnal variations and fluctuations in NO_2 concentration much better compared to an XGBoost model trained with TROPOMI data at a representative example site.

IV. Conclusion

Our results demonstrate that machine learning models built using high-resolution remote sensing data can accurately estimate surface-level NO_2 concentrations. Specifically, we show that tree-based machine learning models trained using TEMPO's high resolution spatial and temporal data can provide accurate estimates of hourly surface-level NO_2 concentrations. Our method combines the benefits of high spatial coverage and resolution provided by satellite-based sensors and the data quality of in situ networks. Our solution can allow for the development of air quality monitoring networks in areas that are not close to the sparse ground-based sensor locations, such as those in rural areas. This can enable individuals to get a better idea of local air quality and make better informed decisions. It can also help inform government policy for allocation of grants and improve monitoring for breaches in environmental regulations.

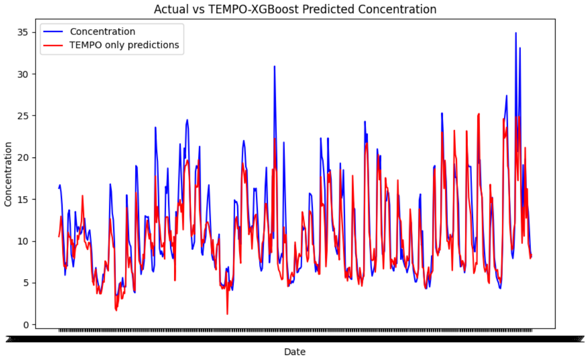

Fig. 6. Hourly Measured NO_2 vs TEMPO-XGBoost model predictions at Sensor ID 840060719004 (San Bernadino, CA) from 5/13 to 7/13

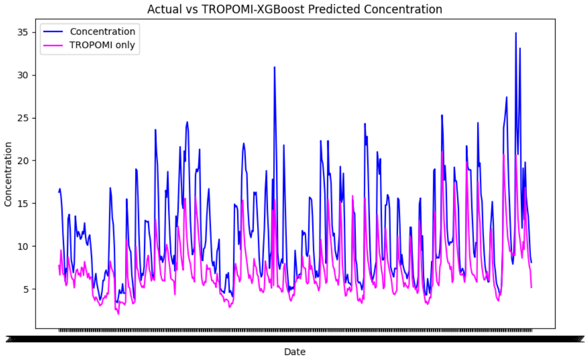

Fig. 7. Hourly Measured NO_2 vs TROPOMI-XGBoost model predictions at Sensor ID 840060719004 (San Bernadino, CA) from 5/13 to 7/13

In the future, we plan to test additional deep learning model architectures such as Long-Short Term Memory (LSTM) models. These new architectures could be leveraged to develop more accurate forecasting models. Additionally, as more data becomes available from the TEMPO sensor, which was just launched last year and started releasing data in May 2024, we expect that the models' performance will improve.

V. Acknowledgment

This research was made possible through the support of the Aspiring Scientists Summer Internship Program (ASSIP) and the support of George Mason University's College of Science, which supports the ASSIP Program.

References

[1] 7 Million Deaths Annually Linked to Air Pollution-WHO Report. 4 Oct. 2017, https://www.unep.org/news-and-stories/press-release/7-million-deaths-annually-linked-air-pollution-who-report.

[2] Air Pollution. https://www.who.int/health-topics/air-pollution.

[3] MJ. Bechle, DB. Millet, B. Dylan, JD. Marshall, "Remote Sensing of Exposure to NO2: Satellite versus Ground-Based Measurement in a Large Urban Area." Atmospheric Environment, vol. 69, Apr. 2013, pp. 345–53. ScienceDirect, https://doi.org/10.1016/j.atmosenv.2012.11.046.

978-1-6654-7345-3/22/$31.00 ©2024 IEEE

[4] JP. Veefkind, et al. “TROPOMI on the ESA Sentinel-5 Precursor: A GMES Mission for Global Observations of the Atmospheric Composition for Climate, Air Quality and Ozone Layer Applications.” Remote Sensing of Environment, vol. 120, May 2012, pp. 70–83. ScienceDirect, https://doi.org/10.1016/j.rse.2011.09.027.

[5] K. Chan, E. Khorsandi, S. Liu, F. Baier, P. Valks, “Estimation of Surface NO2 Concentrations over Germany from TROPOMI Satellite Observations Using a Machine Learning Method.” Remote Sensing, vol. 13, no. 5, Jan. 2021, p. 969. www.mdpi.com, https://doi.org/10.3390/rs13050969.

[6] M. Kim, D. Brunner, G. Kuhlmann, “Importance of Satellite Observations for High-Resolution Mapping of near-Surface NO2 by Machine Learning.” Remote Sensing of Environment, vol. 264, Oct. 2021, p. 112573. ScienceDirect, https://doi.org/10.1016/j.rse.2021.112573.

[7] J. Li, et al. “Comprehensive Evaluations of Diurnal NO2 Measurements during DISCOVER-AQ 2011: Effects of Resolution-Dependent Representation of NOx Emissions.” Atmospheric Chemistry and Physics, vol. 21, no. 14, July 2021, pp. 11133–60. Copernicus Online Journals, https://doi.org/10.5194/acp-21-11133-2021.

[8] https://tempo.si.edu/.

[9] Installation — HARP 1.23 Documentation. https://stcorp.github.io/harp/doc/html/install.html.

[10] H. Setchell, “ECMWF Reanalysis V5.” ECMWF, 19 Feb. 2020, https://www.ecmwf.int/en/forecasts/dataset/ecmwf-reanalysis-v5.

[11] L. Liu, X. Cao, S. Li, N. Jie, “A 31-Year (1990–2020) Global Gridded Population Dataset Generated by Cluster Analysis and Statistical Learning.” Scientific Data, vol. 11, no. 1, Jan. 2024, p. 124. www.nature.com, https://doi.org/10.1038/s41597-024-02913-0.

[12] USGS Science Data Catalog. https://data.usgs.gov/datacatalog/data/USGS:35f9c4d4-b113-4c8d-8691-47c428c29a5b.

[13] Traffic Volumes AADT. https://gisdata-caltrans.opendata.arcgis.com/datasets/d8833219913c44358f2a9a71bda57f76_0/about.

[14] JR. Meijer, M. Huijbregts, K. Schotten, AM. Schipper, “Global Patterns of Current and Future Road Infrastructure.” Environmental Research Letters, vol. 13, no. 6, May 2018, p. 064006. Institute of Physics, https://doi.org/10.1088/1748-9326/aabd42.

[15] AirNow.Gov. https://www.airnow.gov/.